2008 年 6 月 第 35 卷 第 3 期 西安电子科技大学学报(自然科学版) JOURNAL OF XIDIAN UNIVERSITY Jun. 2008 Vol. 35 No. 3

一种提高 DBSCAN 聚类算法质量的新方法

冯少荣[1,2], 肖文俊[1]
(1. 华南理工大学 计算机科学与工程学院,广东 广州 510640;
2. 厦门大学 信息科学与技术学院,福建 厦门 361005)

摘要: 针对基于密度带有"噪声"的空间聚类应用(DBSCAN)聚类算法存在的 3 个主要问题: 输入参数敏感、对内存要求高、数据分布不均匀时影响聚类效果,提出了一种基于遗传方法的 DBSCAN 算法改进方案数据分区中使用遗传思想的 DBSCAN 算法(DPDGA)来提高聚类质量. 利用遗传算法改进 K-means 算法来获取初始聚类中心;对数据进行划分,在此基础上对划分的每一部分使用 DBSCAN 算法进行聚类;合并聚类的结果. 仿真实验表明,新方法较好解决了传统 DBSCAN 聚类算法存在的问题,在聚类效率和聚类效果方面均优于传统 DBSCAN 聚类算法.

关键词: 聚类算法; 遗传算法; 数据划分; 密度
中图分类号: TP301.6 **文献标识码**: A **文章编号**: 1001-2400(2008)03-0523-07

New method to improve DBSCAN clustering algorithm quality

FENG Shao-rong[1,2], *XIAO Wen-jun*[1]
(1. School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, China; 2. College of Information Science and Technology, Xiamen Univ., Xiamen 361005, China)

Abstract: There are three problems along with the Density Based Spatial Clustering of Applications with Noise(DBSCAN) Clustering Algorithm: input sensitivity, desire for too much memory space and the effect of nonuniform data. To solve these problems, a fast Data Partition DBSCAN using Genetic Algorithm(DPDGA) Algorithm is developed which considerably improves the cluster quality. First, the Genetic Algorithm is used to improve the K-means Algorithm to get the initial clustering center. Second, data is partitioned and the DBSCAN Algorithm is applied to cluster partitions. Finally, all clustered result sets are merged. Simulation experiments indicate that the DPDGA Algorithm works well to solve these problems and that both the efficiency and the cluster quality are better than those of the original DBSCAN Algorithm.

Key Words: clustering algorithm; genetic; data partition; density

带有"噪声"的空间聚类应用(DBSCAN)算法[1,2]是基于密度的聚类方法,它要求聚类空间中一定区域(半径 ε)内所包含对象的数目不小于某一给定的阈值 M(最小数目). 并且将密度足够大的那部分记录组成类,它的显著优点是聚类速度快并可以在带有"噪声"的空间数据库中发现任意形状的聚类. 但这个算法使用了 ε 和 M 两个全局变量,需要由用户主观来选择它们,从而影响了最终的聚类结果. 另外,该算法需要把所有数据载入内存,当数据量很庞大时对主存要求较高. 针对 DBSCAN 算法存在的不足,已有一些改进方法[2~5],但效果并不理想. 笔者利用了遗传思想,提出了一种基于遗传方法的 DBSCAN 算法改进方案(DPDGA)来提高聚类质量.

1 DPDGA 算法的基本思想

1.1 DPDGA 算法构造需要考虑的问题

DPDGA 算法的构造应考虑如下 3 个方面的问题.

收稿日期:2007-07-12
基金项目:国家自然科学基金资助(50474033)
作者简介:冯少荣(1964-),男,副教授,华南理工大学博士研究生,E-mail: shaorong@xmu.edu.cn.

1) 若将待处理的数据集按照一定的规则进行划分,则可以降低 DBSCAN 算法对内存的要求和 I/O 要求.此外,通过划分数据,将大数据集划分为多个小数据集,可以减少 DBSCAN 算法构建 R^*-树的时间消耗.

2) 由于在整个的数据集中,数据分布可能是不均匀的,而根据一定规则划分得到的多个局部小数据集,其数据量远小于原始的未划分的数据集.而每个局部小数据集,其数据相对均匀.根据各个局部数据集的情况,选择该局部数据集的参数值,这可以使聚类结果更好.

3) 对数据集的划分,可能将大的聚类划分到两个不同的局部数据集中,也可能将本应属于某个类的点划分到其他的局部数据集中,使该点变成孤立点.因此,需要对局部数据集的聚类结果进行处理,消除划分数据集可能对聚类结果的负面影响.

1.2 DPDGA 算法的主要处理过程

该算法首先采用基于遗传算法[6,7]的方法,获取较优的初始聚类中心;然后根据所获得的初始聚类中心和各个初始聚类中心之间的距离,划分数据集;根据每个数据集的情况,分别选取每个局部数据集的 M_i[8](数据集 i 中包含对象的最小数目)并进行 DBSCAN 聚类;最后,合并各个局部数据集的聚类结果,得到整个数据集的聚类结果.

2 基于遗传算法的聚类中心获取方法

采用 K-means 算法获取聚类中心,对初值具有很强的依赖性,这样的依赖性导致聚类结果的不稳定.针对 K-means 算法初值依赖性问题,目前初始聚类中心的选择方法有以下 8 种:

1) 任意地选取 k 个样本作为初始聚类中心.

2) 凭经验选取有代表性的点作为初始聚类中心.根据个体性质,观察数据结构,筛选出比较合适的样本点.

3) 把全部混合样本直观地分成 k 类,计算各类均值作为初始聚类中心.

4) 通过"密度法"选择代表点作为初始聚类中心.

5) 由 $(k-1)$ 类聚类解出 k 类的代表点.

6) 按最大最小距离聚类法中寻找聚类中心的方法确定初始聚类中心.

7) 进行多次初值选择、聚类,找出一组最优的聚类结果.

8) 采用免疫规划方法进行混合聚类.

除了以上方法外,还有一种扩展的聚类中心选取方法,该方法在类之间有干扰点时效果较好.它与上述方法一个很大的区别是将原来的点延伸到一条线段.这里采用遗传算法获取聚类中心,具体过程如下:

1) 输入种群大小 P_{size},交叉概率 P_c 和变异概率 P_m.

2) 从点集中随机选取点,构成 P_{size} 个个体,用二进制方式进行编码.

3) 对每个个体进行如下操作:a) 进行聚类划分;b) 求当前划分的校正聚类中心;c) 计算适应度.

4) 进行选择、单点交叉和变异操作,记录当前适应度最大的个体 l_{opt}.

5) 如果种群中最优个体连续 35 代无变化,则输出聚类中心,否则转 3)执行.

2.1 编码方式

遗传算法中的进化过程是建立在编码机制基础上的,编码对于算法的性能影响很大.笔者提出了一种新的编码方法,该方法对于一个给定的聚类集合,用 1 表示聚类中心点,0 表示非聚类中心点,这样就可以用一个 0-1 字符串来表示整个聚类集合.采用这种二进制编码方式,结构简单,便于进行交配、变异等操作,对聚类中心的表示明确.

2.2 种群初始化及终止规则

从待处理的点集中随机选 K 个点作为问题的一个解,并进行编码,反复进行 P_{size} (种群大小)次选择.

这里采用的终止条件是:当种群中最优个体连续 35 代无变化时,认为算法已经收敛,适应度最高的个体即为得到的最优聚类中心.

2.3 适应度函数选择

遗传算法在处理过程中以种群中每个个体的适应度值来进行搜索. 每个个体被遗传到下一代中的概率是由该个体的适应度来确定的, 适应度函数对于选择哪两个个体进行交配很重要. 对于每个个体, 采用与 K-means 算法相同的方式进行聚类的划分和重新计算各聚类中心, 然后用每个类中的点与相应聚类中心的距离和作为判断聚类划分质量的准则函数 Δ. Δ 越小, 表示聚类划分的质量越好. Δ 的数学表达式为

$$\Delta(G_1, G_2, \cdots, G_k) = \sum_{i=1}^{k} \sum_{x_j \in G_i} \| x_j - G_i \| .$$

在做遗传操作的过程中, 有可能出现新的个体的中心点的个数不等于用户给定的中心点个数 K, 遗传操作中出现的那些聚类中心个数不等于输入的聚类中心数的新个体, 有可能是真正的最优的个体. 为此, 在这个适度函数里面引入了惩罚函数 $\Phi(K)$, 使得那些真正的最优个体得以保留.

$\Phi(K)$ 函数原形定义为　　$\Phi(K) = 1 + | K_s - K_y |^{1/2}$,

文中采用的适应度函数为　　$F(K) = 1/(\Phi(K)\Delta)$.

2.4 操作算子

选择操作为: 首先将上一代种群中适应度最高的个体保存到下一代种群中, 然后从种群中随机选取 N 个个体$(0 < N < M)$, 将其中适应度最高的个体复制到下一代种群中, 如此重复$(M - 1)$ 次, 就可得到下一代种群的 M 个个体.

交叉操作采用的是单点交叉方法,该方法对于编码长度为 N 的个体, 首先随机生成交叉点位置 c_p $(1 < c_p < N)$, 交换两个父体中位于 c_p 右侧的部分从而生成两个新的子代个体. 交叉操作产生的子代个体除了继承父代个体的信息外, 还会按一定的概率发生变异, 这体现了生物遗传的多样性. 这里使用固定的变异概率 p_m 对子代个体中的二进制位进行变异操作, 变异操作仅对发生变异的位进行"非"运算.

2.5 实验结果及分析

图 1 为模拟数据集. 通过对模拟数据的处理来分析验证基于遗传算法的聚类中心获取方法的有效性,并且与 K-means 算法所得到的最终聚类中心进行比较. 采用的基于遗传算法的聚类中心获取方法的参数如下: 交叉概率 $p_c = 0.9$, 变异概率 $p_m = 0.05$, 种群的大小为 60. 用误差平方和准则函数 J_c 的值对算法得到的聚类中心进行评价, 误差平方和准则函数 J_c 值越小, 表示该组聚类中心选取得越好.

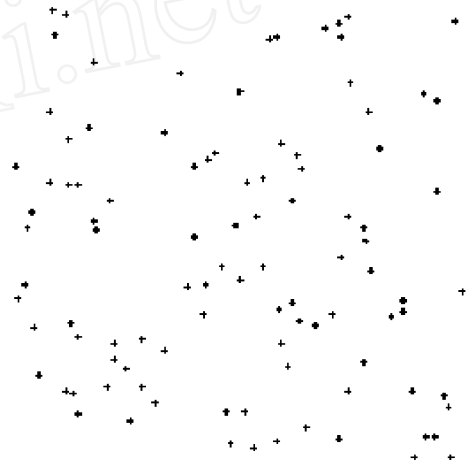

图 1　模拟数据

表 1　模拟数据集的处理结果$(K = 5)$

运行次数	J_c		运行次数	J_c	
	K-means	遗传算法		K-means	遗传算法
1	3595.5	3247.8	4	3534.7	3247.8
2	3248.9	3247.8	5	3289.3	3249.2
3	3545.6	3247.8			

取 $K = 5$ 时, K-means 算法和基于遗传算法的方法对模拟数据集的处理结果在表 1 中给出, 表中第 2 列为使用 K-means 算法得到的准则函数 J_c 的值, 第 3 列为使用基于遗传算法的方法得到的准则函数 J_c 的值. 为了比较, 每个算法分别运行了 5 次, 每次运行的初始解都不一样. 图 2 为其中一次 K-means 算法的聚类结果$(K = 5, J_c = 3595.5)$, 图 3 为其中一次基于遗传算法的方法的处理结果$(K = 5, J_c = 3247.8)$. 两图中均用不同的颜色代表得到的不同的类, 用小"+"表示得到的聚类中心. 由以上结果可以看出, 基于遗传算法的方法得到的准则函数值明显小于 K-means 算法, 即所得到的聚类中心的质量较高.

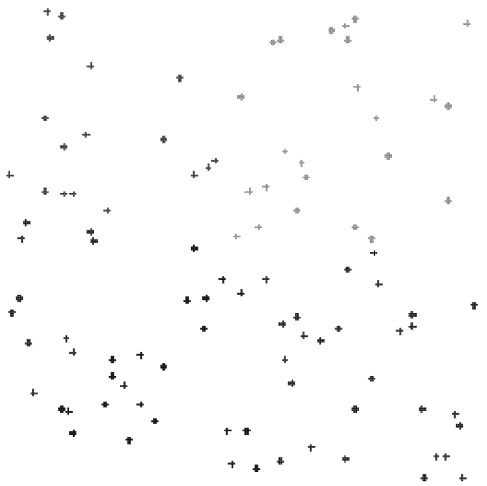

图 2 K-means 聚类结果

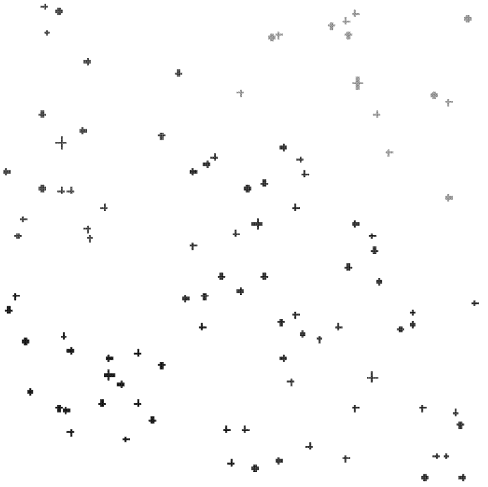

图 3 基于遗传算法的聚类结果

3 数据集的划分

对数据集的划分,涉及到以下定义[8].

定义 1 对于任一初始聚类中心 I_n,它到离它最近的其他初始聚类中心的距离的一半称为该初始聚类中心的划分距离(PartitionDistance).

定义 2 到初始聚类中心距离不大于 PartitionDistance-Eps 的点称为中心点(CorePoint).若一个区域内的所有点都是某一初始聚类中心的中心点,则称该区域为该初始聚类中心的中心区域(CoreRegion).

定义 3 到初始聚类中心的距离大于 PartitionDistance-Eps 且小于 PartitionDistance 的点称为 Eps-中心点(Eps-CorePoint).若一个区域内的所有点都是某一个初始聚类中心的 Eps-中心点,则称之为该初始聚类中心的 Eps-中心区域(Eps-CoreRegion).

定义 4 若某一点既不是中心点,也不是 Eps-中心点,则称该点为非中心点(Non-CorePoint).若一个区域内的所有点都是某一个初始聚类中心的非中心点,则称之为非中心区域(Non-CoreRegion).

整个数据集包括:各个初始聚类中心的中心区域、Eps-中心区域以及非中心区域.一个初始聚类中心的中心区域及其 Eps-中心区域组成一个局部数据集.整个数据集可以被划分为各个初始聚类中心的局部数据集以及各个初始聚类中心共同的非中心区域.由于初始聚类中心的中心区域被该初始聚类中心的 Eps-中心区域环绕,使得每个初始聚类中心的中心点都可以相对独立地进行 DBSCAN 聚类,即各个中心区域的聚类可以并行完成.

局部数据集参数 M_i 的确定 取一个固定的 ε 值,对每个局部数据集分别计算 M_i 的值.每个局部数据集的参数 M_i 为[8]

$$M_i = (E_i / t_i) N_i \quad , \tag{1}$$

其中,N_i 是以 I_i 为中心的局部数据集中数据点的个数;E_i 代表以 I_i 为中心的半径为 ε 的数据集的数据点的超长方体的体积;t_i 代表能够包含所有以 I_i 为中心的局部数据集的数据点的超长方体的体积;而 E_i 根据所属的不同维数,取值分别为

一维数据: $E_i = 2 \times \varepsilon$, (2)

二维数据: $E_i = \pi \times \varepsilon^2$, (3)

三维数据: $E_i = (4/3) \times \pi \times \varepsilon^3$. (4)

对于高维数据,E_i 的值可以近似为:边长为 $2 \times \varepsilon$,包围半径为 ε 的超球体的超立方体的体积.文中主要在二维空间上进行讨论,所采用的 M_i 的计算公式为

$$M_i = (\pi \times \varepsilon^2 / t_i) N_i \quad . \tag{5}$$

4　局部数据集聚类结果的合并

1) 两个类 A 和 B 的合并. 设点 a,b分别为类 A 和类 B 中的点, 若满足以下任一条件, 则合并类 A 和类 B.

①b 从 a 关于 ε 和 M_A 密度可达, 而 a 从 b 关于 ε 和 M_B 密度不可达,则取 $M = M_A$, 合并类 A 和类 B.

②a 从 b 关于 ε 和 M_B 密度可达, 而 b 从 a 关于 ε 和 M_A 密度不可达,则取 $M = M_B$, 合并类 A 和类 B.

③a 从 b 关于 ε 和 M_B 密度可达, 同时 b 从 a 关于 ε 和 M_A 也密度可达,则取 $M = \min\{M_B, M_A\}$, 合并类 A 和类 B.

2) 归并噪音点. 处在分区线附近的噪声点可能是全局中某个类的边界点, 必须考虑将这些临时噪声点归并到相邻分区的某个类中. 一个噪声点 p 被归纳入一个类 C, 当且仅当: ①p 和 C 不处于同一个分区中. ②存在点 q, 点 q 是位于类 C 与类 C 所在的局部数据集 I 的 ε 中心区域的重叠部分的核心对象, 且满足条件 $\mathrm{distance}\{p, q\} \leq \varepsilon$.

5　DPDGA 算法的基本框架

(1) DPDGA 算法框架描述如下:

输入: 控制参数, 聚类数据集.

输出: 聚类结果.

Step 1　初始化控制参数, 包括: 种群大小 P_{size}, 交叉概率 p_c, 变异概率 p_m.

Step 2　从点集中随机选取点, 构成 P_{size} 个个体, 用二进制方式进行编码.

Step 3　对每个个体进行以下操作: ①进行聚类划分. ②求当前划分的校正聚类中心. ③计算适应度.

Step 4　进行选择、单点交叉和变异操作, 记录当前适应度最大的个体 l_{opt}.

Step 5　如果种群中最优个体连续 35 代无变化, 则当前最优个体为所求聚类中心, 否则返回 Step 2.

Step 6　对每个聚类中心, 计算到其他聚类中心距离, 取其中最近距离一半为该初始聚类中心的 PartitionDistance.

Step 7　根据得到的各个初始聚类中心的 PartitionDistance 划分数据集.

Step 8　对于各个局部数据集并行进行以下操作: ①计算各个局部数据集的参数 M_i. ②使用 DBSCAN 算法进行聚类.

Step9　合并各个局部数据集的聚类结果.

Step10　输出聚类结果, 算法结束.

(2) 实验结果

为了检验 DPDGA 算法的可行性和有效性, 对 DBSCAN 算法和 DPDGA 算法进行了对比实验. 实验采用 VC + + 实现, 在 PⅣ2. 4G, 512M 内存的计算机上进行. 为了体现 DPDGA 算法对分布不均匀的数据集的聚类效果, 使用 PB 产生了如图 4 所示的分布不均匀的数据集. 图中左边的部分明显地可以分为两个类, 且数据分布密集, 而右边的部分, 数据分布稀疏.

用基于遗传算法的方法对该数据集进行处理, 得到的初始聚类中心如图 5 所示. 图中, 小“ + ”表示初始聚类中心. 在图 6 所示得到的初始聚类中心的基础上, 使用笔者提出的 DPDGA 算法进行聚类, 得到如图 7 所示的结果. 图中用一种颜色表示一个聚类. 用大的“ + ”表示噪音点.

为了与原始的 DBSCAN 算法的聚类结果进行比较, 图 5 给出了原始的 DBSCAN 算法对模拟数据集的聚类结果.

从两个算法的聚类结果可以看出:原始的 DBSCAN 算法对于不均匀的数据集的聚类质量欠佳, 它把左边相对密集的两个类合并成为了一个类, 同时对右边相对稀疏的数据点, 把本来应该被聚类到某个类中的数据点处理成了噪音. DPDGA 算法由于对数据集进行了划分, 根据各个局部数据集的情况选取不同的参数值, 成功地将左边两个较密集的类分成了两个类, 而不是处理成了一个类. 对于右边较稀疏的数据点也

进行了较符合数据分布情况的聚类,被算法处理成为噪音点的数据点明显少于原始的 DBSCAN 算法,而这也是比较符合数据分布情况的.由此可见,显然改进后的算法得到的聚类结果更能体现数据的分布特征,聚类质量更高.DBSCAN 算法与 DPDGA 算法所需的时间比较如图 8 所示.

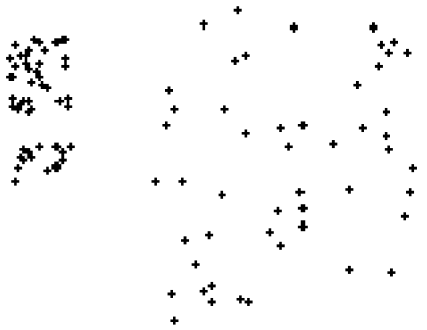

图 4 二维模拟数据集

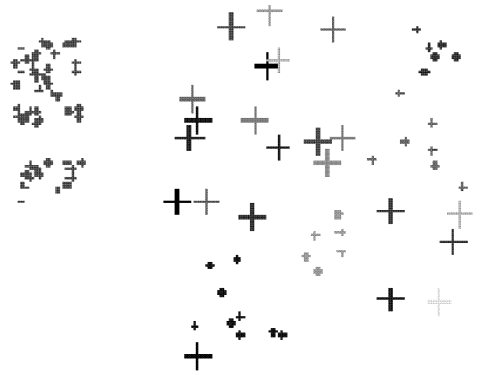

图 5 原始 DBSCAN 算法的聚类结果

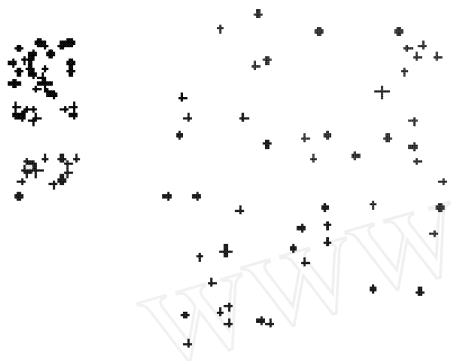

图 6 基于遗传方法得到的初始聚类中心

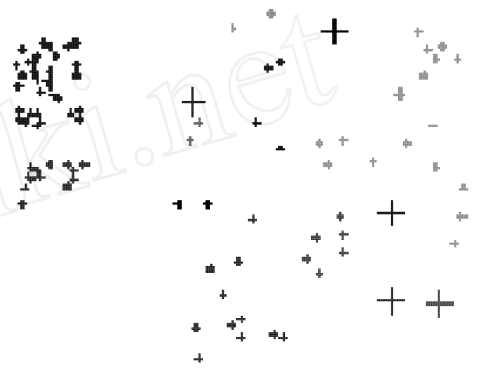

图 7 DPDGA 得到的聚类结果

从图 8 可知,当数据量增大时,DPDGA 算法所需时间的增幅明显比 DBSCAN 算法小,因此,DPDGA 算法比 DBSCAN 算法快.

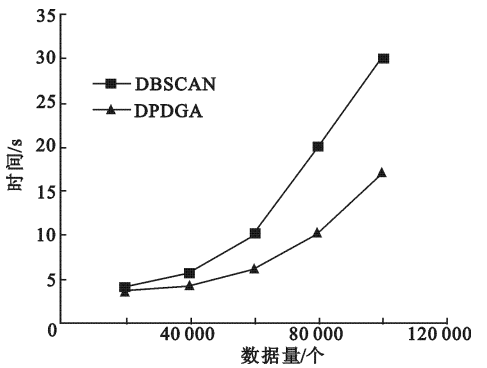

图 8 算法改进前后的运行时间比较

6 结束语

DBSCAN 算法能够在有"噪音"的数据集中发现任意形状的聚类,但当数据量很大时,所要求的内存和 I/O 消耗都较大.当数据分布不均匀时,由于使用统一的全局变量,使得聚类的效果差.笔者针对 DBSCAN 存在的问题,提出了 DPDGA 算法.该算法将数据集划分为多个局部,在对每个局部进行聚类时,选用不同的参数值,最后再合并各个局部的聚类结果.实验结果表明,DPDGA 算法在聚类质量上优于原始的 DBSCAN 算法.

与传统的 DBSCAN 算法相比 DPDGA 算法的优缺点如下:

遗传算法的一个优点是不需要待解决问题领域的特殊知识,DPDGA 算法基于遗传算法获取较优的初始聚类中心,克服了 K-means 算法的初值依赖性.根据获取的初始聚类中心对数据集进行划分,并对划分得到的各个局部数据集使用原始的 DBSCAN 算法聚类,最后合并各个局部数据集的聚类结果.由于不是依赖于人们对待处理数据的先验知识,而是根据各个局部数据集的实际情况,确定算法中的参数值,这样减

少了人为设定参数值的随机性.较好地解决了 DBSCAN 算法输入参数敏感问题.由于划分了数据集,降低了 DBSCAN 算法对内存的要求.并且划分后的每个数据集数据分布相对较均匀,在此基础上使用 DBSCAN 算法进行聚类,可以显著提高聚类的质量.解决了数据分布不均匀影响 DBSCAN 算法聚类效果的问题.

DPDGA 算法还存在以下几个方面需要进一步完善和深入研究:

(1) 选取初始聚类中心时,聚类中心的数目 K仍是人为设置的,不一定符合数据的分布特征. K值的确定和确定 K值方法的评价方式都需要进一步研究.

(2) 针对一个特定的数据集,在确定初始聚类中心时,就基于遗传算法的方法和笔者提到的其他方法,究竟采用哪种方法最佳,有待进一步研究.

(3) DPDGA 算法中,划分数据集是根据得到的初始聚类中心以及各个聚类中心之间的距离.这种划分对于大多数的数据集而言是比较合理的,但是对于一些具有特殊特征的数据集而言,这样的划分是否合适也是今后要探讨的问题.

参考文献:

[1] Ester M, Kriegel H P, Sander J, et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]// Proceedings of the Second International Conference on Knowledge Discovery and Data Mining(KDD-96). Portland, Oregon: AAAI Press, 1996: 226-231.

[2] 蔡颖琨,谢昆青,马修军. 屏蔽了输入参数敏感性的 DBSCAN 改进算法[J]. 北京大学学报(自然科学版),2004,40(3):480-486.
Cai Yingkun, Xie Kunqing, Ma Xiujun. An Improved DBSCAN Algorithm which is Insensitive to Input Parameters[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2004, 40(3): 480-486.

[3] 周水庚,周傲英,曹晶,等. 一种基于密度的快速聚类算法[J]. 计算机研究与发展,2000,37(11):1187-1192.
Zhou Shuigeng, Zhou Aoying, Cao Jing, et al. A Fast Density-based Clustering Algorithm[J]. Journal of Computer Research and Development, 2000, 37(11): 1287-1292.

[4] 周水庚,范晔,周傲英. 基于数据取样的 DBSCAN 算法[J]. 小型微型计算机系统,2000,21(12):1170-1174.
Zhou Shuigeng, Fan Yu, Zhou Aoying. A Sampling-based DBSCAN Algorithm[J]. MINI-MICRO Computer System, 2000, 21(12): 1170-1174.

[5] 周水庚,周傲英,曹晶. 基于数据分区的 DBSCAN 算法[J]. 计算机研究与发展,2000,37(10):1153-1159.
Zhou Shuigeng, Zhou Aoying, Cao Jing. A Data Partitioning Based DBSCAN Algorithm[J]. Journal of Computer Research and Development, 2000, 37(10): 1153-1159.

[6] Lopes F M, Pozo A T R. Genetic Algorithm Restricted by Tabu Lists in Data Mining[C]// 21st International Conference of the Chilean Computer Science Society(SCCC2001). Punta Arenas, Chile: IEEE Computer Society, 2001: 178-185.

[7] Maulik U, Bandyopadhyay S. Genetic Algorithm-Based Clustering Technique[J]. Patten Recognition, 2000, 33(9): 1455-1465.

[8] Dash M, Liu H, Xu X W. 1 + 1 > 2′: Merging Distance and Density Based Clustering[C]// Proceedings of the 7th International Conference on Database Systems for Advanced Applications. Washington: IEEE Computer Society, 2001: 32-39.

(编辑:齐淑娟)